Second-best congestion pricing in general static transportation networks with elastic demands

Verhoef, E.T.

2000

document version
Early version, also known as pre-print

TI 2000-078/3
Tinbergen Institute Discussion Paper

Second-best Congestion Pricing in General Static Transportation Networks with Elastic Demands

Erik T. Verhoef

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Keizersgracht 482
1017 EG Amsterdam
The Netherlands
Tel.: +31.(0)20.5513500
Fax: +31.(0)20.5513555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31.(0)10.4088900
Fax: +31.(0)10.4089031

SECOND-BEST CONGESTION PRICING IN GENERAL STATIC TRANSPORTATION NETWORKS WITH ELASTIC DEMANDS*

Erik T. Verhoef**
Department of Spatial Economics
Free University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
Phone: +31-20-4446094
Fax: +31-20-4446004
E-mail: everhoef@econ.vu.nl
http://www.econ.vu.nl/medewerkers/everhoef/et.html

Key words: congestion, road pricing, networks, second-best
JEL codes: R41, R48, D62

Abstract

This paper studies the second-best problem where not all links of a congested transportation network can be tolled. The second-best tax rule for this problem is derived for general static networks, so that the solution presented is valid for any graph of the network and any set of tolling points available. A number of known second-best tax rules are shown to be special cases of the general solution presented. It is further demonstrated that, for instance by using the concept of 'virtual links', the same method can be applied to a broader class of second-best problems in static networks. Finally, a small network is used to demonstrate numerically that an interior second-best optimum need not always be unique, and need not always exist. However, both examples require extreme differences in marginal external costs across links, and a non-optimal toll-point to be used, which casts doubt on the practical relevance of this complication.

*The author would like to thank Robin Lindsey, Simon Shepherd, Kurt van Dender and two anonymous referees for very useful and inspiring comments on an earlier draft. Any remaining deficiencies, of course, are the author's responsibility alone.

**The author is affiliated to the Tinbergen Institute, Keizersgracht 482, 1017 EG Amsterdam. The research of Erik Verhoef has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

1. Introduction

Pigouvian marginal external cost pricing (Pigou, 1920) is widely accepted among transport economists as the first-best bench-mark solution in the regulation of road transport externalities. It is, however, almost equally commonly recognized that the necessary assumptions for the practical applicability of this standard Pigouvian tax rule will seldom, if ever, be met in reality. These assumptions include, for instance, that optimal charging mechanisms are available, allowing the regulator to set perfectly differentiated taxes for all road users and on all links of the network; that first-best conditions prevail throughout the economic environment to which the transport system under consideration belongs; and that all users and the regulator have perfect information on traffic conditions and tolls.

Indeed, such assumptions are quite unrealistic, and second-best issues in transport policies have accordingly received ample attention in the literature. For instance, Wilson (1983), and d'Ouville and McDonald (1990) study optimal road capacity with sub-optimal congestion pricing; Braid (1989) and Arnott, De Palma and Lindsey (1990) consider uniform or step-wise pricing of a bottleneck; and Arnott (1979) and Sullivan (1983) look at congestion policies through urban land-use strategies. A classic problem in second-best transport regulation concerns the two-route problem, where an untolled alternative road is available parallel to a toll road. This problem has for instance been studied by Lévy-Lambert (1968), Marchand (1968), and more recently also by Braid (1996), Verhoef, Nijkamp and Rietveld (1996), and Liu and McDonald (1999). Glazer and Niskanen (1992) study second-best optimal parking fees for a city centre where through-traffic as well as road users with access to private parking places cannot be charged. Verhoef, Emmerink, Nijkamp and Rietveld (1996) consider second-best congestion tolls under conditions of stochastic congestion and imperfect information. A recurring result in such studies is that second-best tax-rules – set so as to maximize social welfare given the persistence of the second-best distortion – are generally different from the simple Pigouvian rule.

This paper aims to solve the second-best problem where only a subset of links can be tolled for ‘general’ networks, with any possible shape. This type of problem will become increasingly relevant when the foreseen introduction of electronic road charging for a growing number of urban areas becomes reality (Small and Gómez-Ibáñez, 1998). A specific example concerns ‘pay-lanes’, comparable to the two-route problem just mentioned. Another example concerns the determination of optimal cordon charges, for a toll-ring around a city centre. Often, such second-best problems may be ‘self-imposed’ by the regulator, in particular when it is considered inefficient to collect charges on all links of a network, rather than on a subset of links only. This could be the case if relatively high costs are associated with installing the additional tolling equipment, while only relatively low social benefits would be expected to arise from having the additional tolls available. Especially with electronic tolling, such a cost structure may often be the rule rather than the exception.

The analysis considers congestion as the only relevant externality, and concerns a road network. The purpose is to derive the second-best optimal tax rules that would apply for any set of toll-points on any congested transportation network. The analysis pertains to static

networks only, and assumes deterministic equilibria with perfect information. Generalizations to dynamic networks, and to cases with imperfect information and stochasticity, are left as topics for future research. Next, apart from having a possibly different willingness to pay, and possibly different nodes of origin and destination, the (potential) users of the network are assumed to be homogenous.[1] Furthermore, users consider different routes serving the same origin-destination pair as perfect substitutes, so that Wardropian equilibria prevail, with equalized generalized costs for used routes (see Section 2 below).[2]

Three recent related papers can be mentioned. The first is Yan and Lam (1996), who present an algorithm for finding optimal tolls for the same type of problem where not all links can be tolled. In contrast to the present analysis, they consider inelastic demand, and in addition present an algorithm that uses derivatives of link flows with respect to tolls only, and that is therefore not based on the analytical solution of the general second-best problem as presented here. The second is Hearn and Yildirim (1999), who analyze various additional objectives that can be considered, apart from the maximization of welfare, when there is no unique toll vector decentralizing the first-best ('system') optimum. Similar to the present paper, they consider elastic demands and tolls on subsets of links only. However, their analysis concerns alternative first-best pricing schemes, only. Thirdly, May and Milne (2000) consider various second-best tolling schemes in a network model for Cambridge, including cordon, distance-based, time-based and congestion-based charging. They do not, however, consider second-best optimal tolls, but various exogenously determined charge levels instead.

This paper is organized as follows. The next section introduces the notation, and discusses the uniqueness of equilibrium values of some key variables in general transportation networks. Section 3 presents the second-best optimization problem and its solution, and considers the related important question of the optimal location of additional toll-points. Section 4 shows that the general solution obtained indeed is a generalization of earlier results in the literature, and presents some further possible applications of the general model. Section 5 considers the existence and uniqueness of interior second-best optima. Section 6 concludes.

---

[1] At least two types of heterogeneity would be relevant in reality. First, drivers may differ with respect to value of time. Verhoef and Small (1999) consider second-best congestion pricing with heterogeneous drivers on a three-link network, assuming elastic demand and a continuum of values of time. Even for such a small network, they are unable to derive a closed-form expression for a second-best toll on one of two parallel links. Under those conditions, second-best and first-best pricing will lead to a separation of traffic over parallel links by value of time. One of their results is that this increases the relative efficiency of second-best pricing, compared to the case with homogeneous users. Second, drivers may differ with respect to their marginal impact on others' travel times (i.e. trucks or buses *vs* passenger cars). However, the extension to cross-link effects in the Appendix in fact makes it possible to apply the model directly to multiple user groups, by specifying them as occupying copies of the original network (Smith 1979, citing Dafermos 1973). I owe this observation to an anonymous referee.

[2] It is worth pointing out that the methodology proposed could in fact be used directly for the more general problem with imperfect substitutes if 'virtual links' (see also Section 4) could be used to reflect the (apparent) perceived generalized cost difference between alternative routes or modes. A rising marginal private cost function for such (uncongested) virtual links could then represent that individuals differ with respect to the perceived generalized cost difference. If, alternatively, cross demand relations are used to reflect imperfect substitutability, a more fundamental modification of the present framework would be needed, not in the least place because consumer surplus in such cases is not uniquely defined unless additional restrictive assumptions involving symmetrical cross-effects are made (see for instance Liu and McDonald, 1999, p. 162, for an example with two time periods).

2. Notation, equilibria and uniqueness

The analysis in this paper pertains to a general transportation network $\mathcal{G}$ with continuous numbers of users. This network consists of a set of nodes and a set of directed links (arcs). Any pair of distinct nodes can be an origin-destination (OD-)pair, and the demand for trips between such an OD-pair is not restricted to be perfectly inelastic. The notation to be used is presented in Table 1 (where primes denote derivatives).

$\mathcal{N}$	the set of nodes in the network
$\mathcal{I}$	the set of OD-pairs, denoted i=1,…,I or k=1,…,I
N_i	the continuous number of users (or OD-flow) for OD-pair i, with $N_i \geq 0$
$D_i(N_i)$	the inverse demand function for trips for OD-pair i, with $D_i' \leq 0$
$\mathcal{J}$	the set of directed links in the network, denoted j=1,…,J or m=1,…,J
N_j	the continuous number of users (or link-flow) on link j, with $N_j \geq 0$
$c_j(N_j)$	the average cost function for the use of link j, with $c_j' \geq 0$
$\mathcal{P}$	the set of non-cyclical paths in the network, denoted p=1,…,P or q=1,…,P
N_p	the continuous number of users (or path-flow) for path p, with $N_p \geq 0$
$\mathcal{P}_i$	the set of non-cyclical paths for OD-pair i, denoted p_i=1,…,P_i
δ_{jp}	a dummy equal to 1 if link j belong to path p, and to 0 otherwise
δ_j	a dummy equal to 1 if a toll can be charged on link j, and to 0 otherwise
f_j	the level of the toll on link j if δ_j=1
i or k	index for OD-pairs
j or m	index for links
p or q	index for paths
δ_{ip}	a dummy equal to 1 if $p \in \mathcal{P}_i$ and $\sum_{j=1}^{J} \delta_{jp} \cdot \left(c_j(N_j) + \delta_j \cdot f_j\right) - D_i(N_i) \leq 0$, and to 0 otherwise

Table 1. Notation

It is assumed that all relevant functions $D_i(N_i)$ and $c_j(N_j)$ are continuous and smooth. The cost functions represent generalized user costs including monetized time costs, and are upward sloping with congestion. In the analysis below, congestion is assumed to be link-specific. The more general case, where the travel time on a link may also depend on the usage of other links, and where there is nodal congestion, is presented in the Appendix. It turns out to be a straightforward generalization of the analysis presented below. For a dynamic generalization of the present model, for instance based on Vickrey's (1969) model of bottleneck congestion, account should indeed be taken of the possibility that when the arrival rate of users at the tail of a link exceeds its capacity, queuing will occur. This would directly affect the cost levels on upstream links. A static model, however, by definition cannot give a meaningful representation of cases where arrival rates exceed capacities (Verhoef, 1999). The assumption that congestion is link-specific may then often be acceptable, unless intersections are considered to be an important source of congestion (see the Appendix).

Because every path p connects one unique OD-pair, defined by the nodes at the tail of the first arc and the head of the last arc, we have for the total number of paths:

$$P = \sum_{i=1}^{I} P_i \tag{1}$$

Since we are dealing with a static network, the use of a link is defined as:

$$N_j = \sum_{p=1}^{P} \delta_{jp} \cdot N_p \tag{2}$$

An important equilibrium concept is Wardrop's (1952) first principle, stating that for every OD-pair i the costs for used paths must be the same and that there are no unused paths with strictly lower costs. For the general case where the demand functions $D_i(N_i)$ are not necessarily perfectly inelastic, this can be represented with the following complementary slackness equilibrium conditions (see, for instance, Smith, 1979):

$$N_p \geq 0; \quad \sum_{j=1}^{J} \delta_{jp} \cdot \left(c_j + \delta_j \cdot f_j\right) - D_i \geq 0 \quad \text{and} \quad N_p \cdot \left(\sum_{j=1}^{J} \delta_{jp} \cdot \left(c_j + \delta_j \cdot f_j\right) - D_i \right) = 0$$

$$\forall\ p \epsilon \mathcal{P}_i \tag{3}$$

(the arguments in the cost and demand functions are dropped whenever this does not lead to confusion). Compared with the case of inelastic demands, equation (3) therefore adds the economic equilibrium principle that marginal benefits should be equal to marginal private costs to the standard Wardrop condition. The fact that Wardrop's principle allows a formulation of network problems in terms of variational inequalities (VI) has been recognized by for instance Dafermos (1980) and Nagurney (1999).[3] Inspection of (3) reveals that the dummy variable δ_{ip} defined in Table 1 takes on the value of 1 only if path p from the set $\mathcal{P}_i$ is among those that may be used in the equilibrium by travellers between OD-pair i. Such paths with $\delta_{ip}=1$ will be called 'relevant paths' in the sequel. However, for some of the relevant paths, N_p actually still may be equal to zero in the equilibrium, as will become clear when the uniqueness of the various variables in an equilibrium is considered below. First, however, a final identity can be given, equating the usage for a OD-pair to the sum of usage on all relevant paths connecting that OD-pair:

$$N_i = \sum_{p=1}^{P} \delta_{ip} \cdot N_p \tag{4}$$

The uniqueness of equilibria on a network as described above can be defined in various ways, which are not quite equivalent. We consider three measures for uniqueness of an equilibrium for a given vector of tolls (ODF-, LF- and PF-uniqueness). In addition, we distinguish two measures for uniqueness of optimal toll vectors (T-uniqueness and LO-uniqueness).

First, for a given toll-vector **f** (possibly **0**), a unique equilibrium in terms of OD-flows (the vector $\mathbf{N_i}$; *ODF-uniqueness*) and link-flows (the vector $\mathbf{N_j}$; *LF-uniqueness*) can be expected to exist under rather general conditions, in particular if $D_i'(N_i)<0$ and $c_j'(N_j)>0$ for all relevant i and j over the relevant ranges (see, for instance, De Palma and Nesterov, 1998). It is assumed throughout this paper that such a unique solution exists and is stable. However,

[3] An important advantage of VI over alternative formulations is that the existence and uniqueness of a solution can be guaranteed under relatively mild conditions (Nagurney, 1999). For instance, 'asymmetric link-interactions' can easily be accommodated. For the present purpose of finding a second-best optimum in a setting without direct link-interactions, it is unlikely that a formulation in terms of VI would yield additional qualitative insights. However, variational inequality theory could for instance be used to prove existence of a second-best optimum with asymmetric link-interactions, which could occur in the case presented in the Appendix.

this does not imply that the solution will be necessarily unique also in path-flows (the vector $\mathbf{N_p}$; *PF-uniqueness*). This can be illustrated with the simple network shown in Figure 1, where four links (1-4) connect two OD-pairs (I-III and II-III). Both OD-pairs have two paths: in the last part of the trip, either link 3 or 4 can be chosen by both 'I-drivers' and 'II-drivers'. Assume that the regulator can set tolls on each of the four links. Consider the first-best optimum, where the tolls f_j are each set equal to the marginal external congestion costs (mec_j) on these links, and suppose that all OD- and link-flows are positive. This equilibrium is not PF-unique: after interchanging a I-driver on link 3 with a II-driver on link 4, the same equilibrium in OD-flows and link-flows (and hence also in terms of total and marginal benefits and costs) results, although the equilibrium has changed in terms of path-flows.

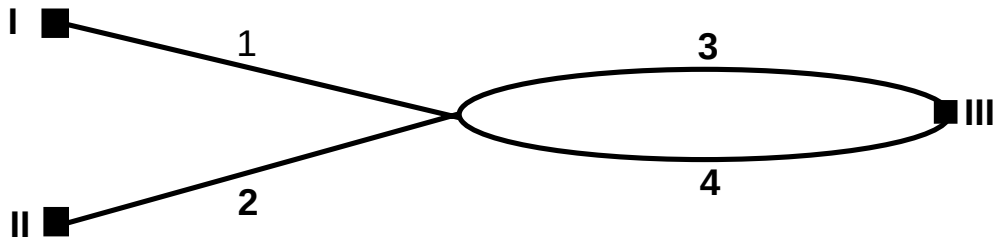


Figure 1. Non-uniqueness of path-flows and link-tolls in a simple network

The same network can be used to illustrate *T-uniqueness*, which refers to the existence of a unique vector $\mathbf{f}$ producing a given first-best or second-best optimum in terms of $\mathbf{N_i}$ and $\mathbf{N_j}$. The first-best optimum in the network in Figure 1 is not T-unique: the same optimum can be realized by adding any constant to f_1 and f_2, and subtracting it from f_3 and f_4. Hence an infinite number of different optimal vectors $\mathbf{f}$ exist. Hearn and Yildirim (1999) consider various additional objectives to select a first-best toll vector when the optimum is not T-unique.

T-uniqueness and PF-uniqueness do not necessarily imply each other. When links 3 and 4 in Figure 1 were merged to one single link, the first-best optimum is PF-unique but still not T-unique. When instead a third OD-pair is added, with the same destination but with the intersection of the four links as origin, the first-best optimum is T unique (with four tolls $f_j=mec_j$) but still not PF-unique. Therefore, the main point is that for general networks, one cannot be sure whether PF-uniqueness or T-uniqueness will prevail. This, in turn, will be reflected in the general solution to be derived below.

Apart from T-uniqueness, however, it is also possible that multiple local second-best optimal toll vectors exist that not only differ in terms of the vectors $\mathbf{f}$, but also in terms of the vectors $\mathbf{N_i}$ and $\mathbf{N_j}$. We will refer to this type of uniqueness as *LO-uniqueness* (local optimum uniqueness), to emphasize that when LO-uniqueness fails, different second-best optima will have different OD-flows, link-flows, and – generally – welfare levels. Section 5 will demonstrate that with second-best pricing, LO-uniqueness cannot be guaranteed. Note that T-uniqueness and LO-uniqueness are two entirely different concepts, and that there is again no reason to expect that the either one should imply the other.

3. Solving the second-best optimization problem

The stage is now set to derive the second-best optimal congestion tolls when tolls can be charged only on a given subset of links. As a matter of fact, the first-best problem where tolls can be charged on all links is, of course, a special case of this general second-best problem. It is assumed that, given the constraint, the regulator sets tolls so as to maximize social welfare, defined as total benefits minus total costs. Benefits are determined according to the Marshallian measure. The formal maximization problem can thus be stated as follows:

$$\underset{f_j \;\forall\; j \text{ with } \delta_j=1}{\text{MAX}} W = \sum_{i=1}^{I} \int_0^{N_i} D_i(x_i)dx_i - \sum_{j=1}^{J} N_j \cdot c_j(N_j) \quad \text{s.t. (3)} \;\forall\; i \tag{5}$$

This maximization problem is in fact a bi-level optimization problem, where the regulator maximizes an objective, given that the road users maximize their own objective as represented by the Wardropian conditions in (3). Because road users are assumed to treat tolls as given and to be independent of their own behaviour, the problem is in fact a Stackelberg game, where the regulator is the leader and all road users are the followers.

The form of the Wardropian conditions in (3) prevents us from using them directly as constraints in a standard non-linear programming problem (NLP) equivalent with (5). However, it will turn out that we can write (3) as a standard NLP for *interior second-best optima* (as opposed to *corner solutions*), which are defined as optima for which the set of relevant paths does not change due to marginal changes in any of the tolls available. It is tempting to entirely discard corner solutions on the grounds that, in continuous space, there would be a zero probability of finding a second-best optimum where one of the paths is exactly balancing between relevance and irrelevance. However, Section 5 makes clear that this is not a valid argument: a second-best optimum may actually be a corner solution if for relative low values of one of the tolls, the then relevant paths would require a relatively high value of that toll, while with a relatively high value of that toll, the then relevant paths – a different set – would instead require a low toll. Under extreme conditions, and when the regulator has chosen particularly unattractive links as toll points, this may happen. However, given the hypothetical character of this type of optimum, we proceed by considering interior second-best solutions only, which can be expected to be the rule rather than the exception.

For a local interior second-best optimum, the necessary first-order conditions (foc's) can then be found by treating all δ_{ip} as constants (recall that foc's involve marginal evaluation in the optimum only). All irrelevant paths can be discarded, and the Wardropian conditions for relevant paths become equalities. The maximization problem in (5) can therefore be represented by the following Lagrangian :

$$\begin{aligned}\Lambda = &\sum_{i=1}^{I} \int_0^{\sum_{p=1}^{P} \delta_{ip} \cdot N_p} D_i(x_i)dx_i - \sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{p=1}^{P} \delta_{jp} \cdot \delta_{ip} \cdot N_p \cdot c_j\left(\sum_{k=1}^{I}\sum_{q=1}^{P} \delta_{jq} \cdot \delta_{kq} \cdot N_q\right) \\ &+ \sum_{i=1}^{I}\sum_{p=1}^{P} \delta_{ip} \cdot \lambda_p \cdot \left[\sum_{j=1}^{J} \delta_{jp} \cdot \left(c_j\left(\sum_{k=1}^{I}\sum_{q=1}^{P} \delta_{jq} \cdot \delta_{kq} \cdot N_q\right) + \delta_j \cdot f_j\right) - D_i\left(\sum_{q=1}^{P} \delta_{iq} \cdot N_q\right)\right]\end{aligned} \tag{6}$$

where for the transparency of the derivation, but at the expense of some extra notational clutter in particular for the arguments of cost functions, (2) and (4) are substituted into (5).

The first set of terms represent total benefits, summed over all OD-pairs; note that the total OD-flow is determined according to (4). The second set of terms represent total costs, summed over all links in the network; note that the total link-flow is determined according to (2). The third set of terms represent the constraints caused by the equilibrium conditions that for each relevant path, the marginal benefits will be equal to the average costs plus the fees incurred on the links making up that path. Note that these constraints are consistent with (3) for relevant paths, and that λ_p denotes the Lagrangian multiplier associated with the constraint for path p. These multipliers will be discussed in further detail below. Finally, it ought to be noted that the use of the dummies δ_{ip} (or δ_{iq}) automatically secures that in the determination of the necessary first-order conditions for a local optimum only the relevant paths – which either are used or could be used in the second-best equilibrium – are considered.

The following necessary foc's for an interior second-best optimum can now be derived (arguments in demand and cost functions are again dropped for notational convenience):

$$\begin{aligned}\frac{\partial \Lambda}{\partial N_p} &= \sum_{i=1}^{I} \delta_{ip} \cdot D_i - \sum_{j=1}^{J} \delta_{jp} \cdot \left(c_j + \sum_{k=1}^{I} \sum_{q=1}^{P} \delta_{jq} \cdot \delta_{kq} \cdot N_q \cdot c_j' \right) \\ &+ \sum_{k=1}^{I} \sum_{q=1}^{P} \delta_{kq} \cdot \lambda_q \cdot \left(\sum_{j=1}^{J} \delta_{jp} \cdot \delta_{jq} \cdot c_j' \right) - \sum_{i=1}^{I} \sum_{q=1}^{P} \delta_{ip} \cdot \delta_{iq} \cdot \lambda_q \cdot D_i' = 0 \quad \forall \ p \text{ with } \delta_{ip} = 1\end{aligned} \tag{7}$$

$$\frac{\partial \Lambda}{\partial f_j} = \sum_{i=1}^{I} \sum_{p=1}^{P} \delta_{ip} \cdot \delta_{jp} \cdot \lambda_p = 0 \quad \forall \ j \text{ with } \delta_j = 1 \tag{8}$$

$$\frac{\partial \Lambda}{\partial \lambda_p} = \sum_{j=1}^{J} \delta_{jp} \cdot \left(c_j + \delta_j \cdot f_j \right) - \sum_{i=1}^{I} \delta_{ip} \cdot D_i = 0 \quad \forall \ p \text{ with } \delta_{ip} = 1 \tag{9}$$

Although it was explained above that an ODF- and LF-unique second-best optimum may not be PF-unique, equations (7) show that the foc's with respect to path-flows are used to solve the problem. Path-flows give the necessary connection between the benefit side (in terms of OD-flows) and the cost side (in terms of link-flows) in the model. It may in particular be noted that for a given equilibrium in terms of OD- and link-flows, the value of the right-hand side of (7) is independent of the specific distribution of users from a given OD-pair over the various possible paths, as long of course as the equilibrium conditions shown in equation (3) hold, since the relevant terms only depend on either OD-flows or link-flows.

Substitution of (9) into (7) for each p for which $\delta_{ip}=1$ subsequently yields the following expression for the Lagrangian multipliers λ_p in an equilibrium:

$$\begin{aligned}\lambda_p = \frac{\sum_{j=1}^{J} \delta_{jp} \cdot \left(\sum_{q=1}^{P} \delta_{jq} \cdot N_q \cdot c_j' \right) - \sum_{q=1, q \neq p}^{P} \lambda_q \cdot \left(\sum_{j=1}^{J} \delta_{jp} \cdot \delta_{jq} \cdot c_j' \right)}{\sum_{j=1}^{J} \delta_{jp} \cdot c_j' - \sum_{i=1}^{I} \delta_{ip} \cdot D_i'} \\ + \sum_{i=1}^{I} \sum_{q=1, q \neq p}^{P} \delta_{ip} \cdot \delta_{iq} \cdot \lambda_q \cdot D_i' - \sum_{j=1}^{J} \delta_{jp} \cdot \delta_j \cdot f_j \\ \hline \sum_{j=1}^{J} \delta_{jp} \cdot c_j' - \sum_{i=1}^{I} \delta_{ip} \cdot D_i'\end{aligned} \tag{10}$$

$$\forall \ p \text{ with } \delta_{ip} = 1 \quad \text{and} \quad \forall \ q \text{ with } \delta_{iq} = 1$$

These Lagrangian multipliers, when being unequal to zero, cause the second-best solution to be inferior to the first-best case where tolls can be set on all links. The fact that these multipliers would be zero in the first-best case can be verified by rewriting (6) as if path-tolls f_p could be charged for all paths. This would yield, instead of (8), $\lambda_p=0$ for all relevant paths, and path tolls equal to the sum of marginal external congestion costs on all links used in that path (given by the first of the four terms in the numerator of (10)). This, in turn, can be realized with link tolls each equal to the marginal external congestion costs for that link.

The Lagrangian multipliers λ_p can thus be interpreted as the 'shadow price of non-optimal pricing' in the second-best optimum – which in fact follows directly from the specification of the Lagrangian (6). Although for a general network, no closed-form analytical solution exists with each relevant λ_p independent of the other relevant λ_p's (or λ_q's as they are labelled in (10)), it can be noted that equations (10) will make up a system of X equations, generally linearly independent, in X unkowns (the λ_p's), where X denotes the number of relevant paths in the second-best optimum. Hence, for a given equilibrium in terms of OD- and link-flows, each of these multipliers can be solved for, independent of the value of the other multipliers. The reason that no general analytical solution can be given is, of course, that the expression will depend on the specific network, the tolling points, and the relevant paths.

Further inspection of (10) allows the identification of the terms affecting the size of the 'shadow price of non-optimal pricing' for a relevant path in the second-best optimum. Focusing first on the first and last term in the numerator, this shadow price appears to be increasing in the extent to which the marginal external congestion costs caused (the first term) exceeds the sum of total tolls paid during the trip (the third term). This is conform intuition.

The second and third terms show that, because we are in a second-best optimum, also indirect effects count. The second term shows that the shadow price is decreasing in the extent to which the presence of users from the path considered prevents users from other non-optimally priced relevant paths to use the network. The associated term is in the first place increasing in the shadow prices for these other paths. This reflects that the 'reward' (that is, the reduction in the own shadow price) becomes larger, the larger the shadow prices for the other affected groups are. The term is also increasing in the slope of the average cost functions on the relevant links in the second-best optimum, which reflects that this effect becomes more important as the cost levels on these links are more strongly dependent on link usage. Note that λ_p may have either sign – which was in fact already implied by (8).[4]

The third term shows that the shadow price is also decreasing in the shadow prices for paths that belong to the same origin-destination pair. The distortion due to tolls falling short of marginal external congestion costs for a path are less harmful to efficiency if the implied excess use of that path leads to a lower use of other paths, which too are inefficiently priced. The multiplication by the slope of the demand curve shows that this effect becomes more important when the overall demand for that OD-pair is less price-sensitive. Price changes on one path will then more strongly affect the use of other paths for that OD-pair.

---

[4] This is also the reason for using the dummy variables δ_{ip} and writing the problem as a Lagrangian, instead of using a Kuhn-Tucker formulation where the resulting multipliers would be restricted to be positively signed.

Finally, the denominator of (10) shows that the 'shadow price of non-optimal pricing' for a specific relevant path is decreasing in the sensitivity of the path flow to distorted prices in the second-best optimum. If either the demand for the OD-pair or the 'supply' for the path (represented by the link-cost functions) is fully inelastic, the multiplier vanishes.

As (8) shows, an important feature of the second-best optimum is that the sum of the relevant λ_p's be minimized, which rather intuitively reflects the goal of minimizing the overall distortions due to imperfect pricing. Substitution of (10) into (8) gives the following expression for the second-best optimal congestion fees:

$$f_j = \frac{\sum_{p=1}^{P} \delta_{jp} \cdot \dfrac{\sum_{m=1}^{J} \delta_{mp} \cdot \left(\sum_{q=1}^{P} \delta_{mq} \cdot N_q \cdot c'_m \right) - \sum_{q=1, q \neq p}^{P} \lambda_q \cdot \left(\sum_{m=1}^{J} \delta_{mp} \cdot \delta_{mq} \cdot c'_m \right) + \sum_{i=1}^{I} \sum_{q=1, q \neq p}^{P} \delta_{ip} \cdot \delta_{iq} \cdot \lambda_q \cdot D'_i - \sum_{m=1, m \neq j}^{J} \delta_{mp} \cdot \delta_m \cdot f_m}{\sum_{m=1}^{J} \delta_{mp} \cdot c'_m - \sum_{i=1}^{I} \delta_{ip} \cdot D'_i}}{\sum_{p=1}^{P} \dfrac{\delta_{jp}}{\sum_{m=1}^{J} \delta_{mp} \cdot c'_m - \sum_{i=1}^{I} \delta_{ip} \cdot D'_i}} \qquad (11)$$

$\forall\ j$ with $\delta_j = 1$ and $\forall\ p$ with $\delta_{ip} = 1$ and $\forall\ q$ with $\delta_{iq} = 1$

where, for notational reasons, the index m, when used, denotes links. After the discussion of (10), the interpretation of (11) is actually more easy to give than may seem at first sight. First, the first term $\Sigma_p \delta_{jp}$ in the numerator of (11) shows that only the relevant paths using the link j should be considered directly in the determination of the second-best toll f_j – although, via the terms λ_q in the numerator's numerator, other relevant paths may of course indirectly affect the level of f_j. As a matter of fact, the numerator's numerator again gives the difference between an OD-flow's 'generalized marginal external costs' (corrected for the indirect effect on the usage by other relevant paths) and the total tolls (but now net of the specific toll f_j itself), closely resembling the term already encountered in the numerator of (10). Finally, the further structure of (11) shows that the second-best optimal toll on a link should be a weighted average of the sum of the generalized marginal external costs, minus the tolls paid on other links, for the relevant path-flows. The weights are increasing with the sensitivity of the path-flow to prices in the second-best optimum. This effect reflects what was already observed in the discussion of equation (10).

The fourth term in the numerator's numerator in (11) indicates that second-best tolls need not be T-unique. Likewise, PF-uniqueness need not prevail, for instance when the set of same parallel links are used by different OD-pairs. An important question, however, is whether the first-order conditions (7)-(9) and the implied second-best values of the relevant λ_p's and f_j's imply a unique local second-best optimum in terms of OD-flows and link-flows (assuming that the second-order conditions are fulfilled), which we denoted LO-uniqueness. Section 5 will give examples where LO-uniqueness holds, and where it fails. A general answer to this question therefore cannot be given, as it depends on the exact shape of the network, the selected tolling points, and the shape of the demand and cost functions.

Moreover, Section 5 will also present an example where an interior solution to the second-best problem does not exist, and only a corner solution exists, so that the first-order conditions (7)-(9) and expressions (10) and (11) derived from them do not hold in the second-best optimum. One has to be modest, therefore, and it should be emphasized that the tax rules implied by (11) give necessary conditions for an interior local second-best optimum in a general transportation network only, rather than necessary and sufficient conditions for a global optimum. Under quite general circumstances, however, one would expect only few (if more than one) second-best equilibria supported by taxes as given in (11) to exist, and considering only those equilibria where such taxes apply will generally greatly reduce the task of finding the second-best optimum for a given problem.[5]

Next, second-best tolls as in (11) are not necessarily positive; Section 4.2 and 5.1 present simple cases with negative second-best tolls. In such cases, adding a non-negativity constraint on tolls would reduce welfare. One might object against negative second-best tolls that users might be invited to make cycles, if the network allows so. This, however, would ignore that the second-best tolls (11) are valid when evaluated in the second-best optimal equilibrium only. Unlimited cycling would make the congestion on that particular link so high that the second-best toll would no longer be negative in equilibrium – which would take away the incentive for cycling. When cycling is not worthwhile, negative tolls may be second-best optimal as such. If cycling might occur, we may in fact end up in a corner solution where a negative toll is set such that cycling is just prevented. Therefore, if a second-best optimum found has some negative tolls and was derived considering non-cyclical paths only, it is important to check whether cycling indeed is not worthwhile.

Finally, an important question that is closely related to the above analysis concerns the optimal location of additional toll-points. This question will be relevant not only when an existing tolling system can be extended to cover a larger part of the network, but also when an entirely new tolling system can be installed. Clearly, the most certain way of selecting the optimal location for a single next toll-point would be to calculate the level of welfare under second-best tolling according to (11), for having a toll added on each possible, as yet untolled link. The optimal next toll-point is the one yielding the highest welfare improvement (which could be compared with the costs of adding the toll-point to guarantee a worthwhile investment). For larger networks, however, this procedure may require many calculations, in particular if the problem is somewhat more general, and the optimal location for a (possibly also optimized) number of additional toll-points should be determined.

A general solution to the problem of selecting the optimal toll points – other than the general procedure described in the previous paragraph – probably cannot be given for general networks, due to the discreteness of the problem and the dependence on the prevailing

[5] A general numerical procedure for finding such a second-best optimum in a given transport network model could be based on the following sequence: (step 1) start with zero tolls and calculate the equilibrium; (step 2) calculate the out-of-equilibrium values of the λ_p's for all relevant paths for this equilibrium according to (10); (step 3) calculate the implied out-of-equilibrium tolls for the relevant toll points according to (11); (step 4) apply these tolls by adding them to the (perceived) link costs and calculating the new equilibrium; (step 5) check for convergence and go back to (step 2) if the system has not yet converged. The performance of a comparable algorithm is tested in Verhoef (2000) for a 10-link network, with generally promising results.

network structure. Still, one might hypothesize that the evaluation of (10) or (11) in a given equilibrium, outside a second-best optimum, might provide an efficient means of predicting that link or those links for which it is most efficient to add the next toll-point(s).

In particular, the solution of the system of equations following from subtraction of (9) from (7) for all relevant paths (yielding a number of equations equal to the number of relevant paths in the existing equilibrium) – treating all N_i, N_j, and f_j for links already tolled, as given – would yield 'shadow multipliers' L_p for all relevant paths, consistent with (10), evaluated outside a second-best optimum. The evaluation of the right-hand side of (8) for each untolled link, with L_p substituted for λ_p for all relevant paths, would then suggest the link for which $|\partial\Lambda/\partial f_j|$ is maximized (in the existing equilibrium) as the one for which a marginal change in the zero toll level gives the highest net social benefits.

Likewise, 'shadow tolls' ϕ_j can be calculated in an existing equilibrium by solving the same system of equations with however (8) added, which should yield shadow tolls consistent with (11) evaluated outside a second-best optimum. The link for which $|\phi_j|$ is maximized can be identified as the one for which the weighted generalized marginal external costs have the highest absolute value.

The links suggested by either of these procedures may often be those for which in the new second-best equilibrium, with a toll added on that link, the total social welfare improvement is indeed relatively large.[6] Link-selection procedures based on these indicators may offer a quick and reasonably accurate manner to select the optimal location for a next toll-point, or – by applying it sequentially and including feedback mechanisms that allow one to also remove toll points selected earlier – a number of toll-points. Such hypotheses are tested in Verhoef (2000) for a 10-link network, with encouraging results. Nevertheless, one cannot be sure whether such procedures will always be optimal.

4. Comparing the general solution with earlier results

It is instructive to validate the tax-rule (11) by comparing it to results reported earlier in the literature for second-best problems that are special cases of the general problem considered here. Three such cases will be considered: first-best tolling, the standard two-route problem, and the parking problem studied by Glazer and Niskanen (1992). The section concludes with some thoughts on further possible applications of the general model presented above.

4.1. First-best tolling

The most straightforward special case of the general second-best problem discussed is in fact the first-best problem where tolls can be set on all links. For that case, one would expect

[6] Note in particular that adding a toll point on a link for which $\phi_j=0$ and $|\partial\Lambda/\partial f_j|=0$ will yield no welfare improvement at all, since the optimal toll for this link will then also be $f_j=0$, and the same second-best equilibrium will necessarily result. The most important difference between the two indicators is that ϕ_j focuses on the current degree of under-pricing, but ignores the real impact that a toll would have on traffic flows (responsiveness to pricing is absent), whereas $\partial\Lambda/\partial f_j$ only considers the impact of a marginal change of the as yet non-existing toll, ignoring the question of whether this toll will be high. It may well be that in practical applications, a weighted average or multiplication of both indicators would perform best (see Verhoef, 2000).

equation (11) to be consistent with the simple Pigouvian rule equating the tax to the marginal external congestion costs for each link, as given in (12):

$$f_j = \sum_{p=1}^{P} \delta_{jp} \cdot N_p \cdot c'_j \tag{12}$$

To show that this indeed is the case, first observe that – as was argued before – al λ_p's are equal to zero in the first-best case. Equation (11) then reduces to an expression stating that the toll on a link should be equal to the weighted average (over all relevant paths using j) of the difference between the total marginal external congestion costs for that path (over the entire trip, so over all links including link j itself) minus the tolls paid on all links other than j itself. Evidently, the simple Pigouvian tax in (12) is consistent with this rule (as pointed out in Section 2, for many networks the first-best solution needs not be T-unique).

4.2. The standard two-route problem

The standard two-route problem concerns the network in Figure 2, where an untolled route (U) and a parallel tolled route (T) connect the same single origin-destination pair (I-II). As mentioned, this problem has been studied by Lévy-Lambert (1968), Marchand (1968), Braid (1996), Verhoef, Nijkamp and Rietveld (1996), and Liu and McDonald (1999). These studies have shown that the optimal second-best one-route toll for route T can be written as:

$$f_T = N_T \cdot c'_T - N_U \cdot c'_U \cdot \frac{-D'}{c'_U - D'} \tag{13}$$

Equation (13) shows that this toll should be equal to the marginal external congestion costs on the tolled route minus a term consisting of a fraction (between 0 and 1) of the marginal external congestion costs on the untolled route. Note that (13) may imply a zero or a negative second-best optimal toll.

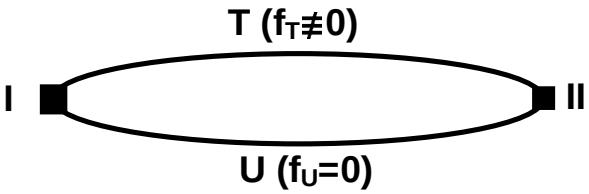


Figure 2. The standard two-route problem

For a further interpretation of (13), see for instance Verhoef, Nijkamp and Rietveld (1996); the important question now is whether (13) is a special case of the general tax rule in (11). To demonstrate that this is the case, it is sufficient to observe that for this problem, the two paths coincide with the two links, so that the necessary first-order conditions (7)-(9) become:[7]

$$\frac{\partial \Lambda}{\partial N_T} = D - c_T - N_T \cdot c'_T + \lambda_T \cdot c'_T - (\lambda_T + \lambda_U) \cdot D' = 0 \tag{14a}$$

[7] These first-order conditions differ from those in Verhoef, Nijkamp and Rietveld (1996) in that the sign of the constraints and hence of the Lagrangian multipliers are now opposite to those in the original formulation. This, of course, does not affect the result.

$$\frac{\partial \Lambda}{\partial N_U} = D - c_U - N_U \cdot c_U' + \lambda_U \cdot c_U' - (\lambda_T + \lambda_U) \cdot D' = 0 \tag{14b}$$

$$\frac{\partial \Lambda}{\partial f_T} = \lambda_T = 0 \tag{15}$$

$$\frac{\partial \Lambda}{\partial \lambda_T} = c_T + f_T - D = 0 \tag{16a}$$

$$\frac{\partial \Lambda}{\partial \lambda_U} = c_U - D = 0 \tag{16b}$$

Fully consistent with (10), using (15), (16b) and (14b) λ_U can then be solved as:

$$\lambda_U = \frac{N_U \cdot c_U'}{c_U' - D'} \tag{17}$$

and substitution of (15), (16a) and (17) into (14a) gives the desired result given in (13).

4.3. Parking policies

A third example concerns the problem of optimal parking fees for congestion management in the case where a subset of road users do not have to pay this fee, for instance because they have access to private parking places (Glazer and Niskanen, 1992). By adding two 'virtual links' to the original one-link network, the problem can be represented as a network problem that allows using the methodology presented in the previous sections.

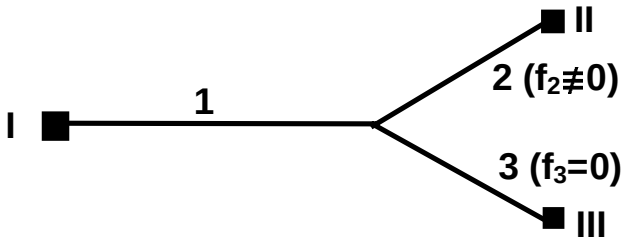


Figure 3. The Glazer and Niskanen (1992) parking problem in a network representation

Figure 3 shows the resulting three-link network, connecting two OD-pairs I-II and I-III, where II denotes priced parking space and III free parking space. Congestion only occurs on the shared link 1. The virtual links 2 and 3 are costless, but a (parking) fee can be charged on link 2. The two paths coincide with the two OD-pairs, so that the necessary first-order conditions (7)-(9) now become (where a subscript II or III denotes the destination; note that $c_2=c'_2=c_3=c'_3=0$):

$$\frac{\partial \Lambda}{\partial N_{II}} = D_{II} - c_1 - N_1 \cdot c_1' + (\lambda_{II} + \lambda_{III}) \cdot c_1' - \lambda_{II} \cdot D_{II}' = 0 \tag{18a}$$

$$\frac{\partial \Lambda}{\partial N_{III}} = D_{III} - c_1 - N_1 \cdot c_1' + (\lambda_{II} + \lambda_{III}) \cdot c_1' - \lambda_{III} \cdot D_{III}' = 0 \tag{18b}$$

$$\frac{\partial \Lambda}{\partial f_2} = \lambda_{II} = 0 \tag{19}$$

$$\frac{\partial \Lambda}{\partial \lambda_{II}} = c_1 + f_2 - D_{II} = 0 \tag{20a}$$

$$\frac{\partial \Lambda}{\partial \lambda_{III}} = c_1 - D_{III} = 0 \tag{20b}$$

where $N_1=N_{II}+N_{III}$. Fully consistent with (10), using (19), (20b) and (18b) λ_{III} can then be solved as:

$$\lambda_{III} = \frac{N_1 \cdot c_1'}{c_1' - D_{III}'} \tag{21}$$

and substitution of (19), (20a) and (21) into (18a) gives:

$$f_2 = N_1 \cdot c_1' \cdot \frac{-D_{III}'}{c_1' - D_{III}'} \tag{22}$$

which is the same as the second-best congestion charge derived by Glazer and Niskanen (1992) (their equation (18)), and is again a special case of (11). This toll is a fraction of the marginal external congestion costs on the real link, where the fraction depends on the demand elasticity for the untolled users (see Glazer and Niskanen, 1992, and Verhoef, Nijkamp and Rietveld, 1995, for further discussions).

4.4. Some further possible applications of the general model

The use of the concept of virtual links in the final case above in fact demonstrates how easily the general model in equations (6)-(11) can be adapted to allow consideration also of different types of second-best policies in a network environment, other than the pure problem caused by the joint existence of tolled and untolled real ('physical') links in the network.

Consider, for instance, the use of peak-hour permits. With such a policy, road users would have to purchase a permit before they are allowed to use the road. However, once they do have such a permit, there would be no further restriction on the use of the network. To determine the second-best optimal price for such permits for a given road network, the regulator therefore has to solve the second-best problem that is caused by the fact that the same single 'toll' (that is: the price of the permit) applies for drivers using different paths, and hence generally causing different levels of marginal external costs. The only adaptation that needs to be made to solve this particular problem using the general network model presented in equations (6)-(11), is to add one single virtual link, with zero costs, on which the regulator can set a toll. This virtual link should be added to all paths, and the optimal toll can then be derived directly according to (11).[8]

Another extension that can be mentioned is the use of distance-based tolls – one of the schemes considered by May and Milne (2000). If the regulator can set only a toll level per vehicle-kilometre travelled, while marginal external congestion costs per vehicle-kilometre

[8] If the regulator wants to use a system of tradeable peak-hour permits, according to the same principles but distributed initially for free, in fact exactly the same problem as described in the main text has to be solved. The second-best optimal number of permits to be issued will then be equal to the number of trips made in the second-best optimum considered in the main text; and the equilibrium price of the permits will be equal to the second-best toll. This holds true, of course, only under the assumption of zero transaction costs.

vary over the network, another second-best problem results. To solve this problem using the proposed model, rewrite the original tolls f_j as $l_j \cdot f$, where l_j denotes the length of link j. Observe that the original first-order condition (8) is then replaced by:

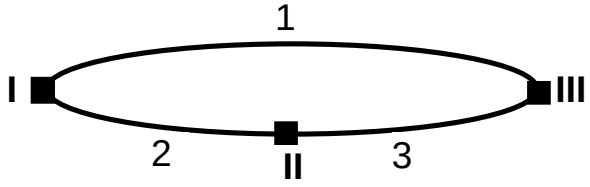
$$\frac{\partial \Lambda}{\partial f} = \sum_{i=1}^{I} \sum_{p=1}^{P} \delta_{ip} \cdot \lambda_p \cdot \sum_{j=1}^{J} \delta_{jp} \cdot \delta_j \cdot l_j = 0 \tag{23}$$

(Note that the formulation still allows cases where the toll is not charged on all links. This could be relevant when the scheme applies to a certain area only, and some users originate from outside this area). In the second-best optimum, now the *weighted* sum of the λ_p's should be equal to zero, where the weight for a path is proportional to the number of tolled kilometres it has. The problem can then be solved analogous to the discussion in Section 3.

Clearly, the two second-best problems just mentioned do not have simple analytical solutions. However, the reason for mentioning these problems here was merely to illustrate that the network model presented in Sections 2 and 3 can easily be extended to deal also with different classes of second-best problems in static transportation networks, other than the pure problem caused by the joint existence of tolled and untolled real links in a network.

5. Existence and uniqueness of interior second-best optima

The first-order conditions (7)-(9), when satisfied, and the implied second-best tolls as given in (11) define an interior second-best optimum. Important questions are whether such interior optima exist, and if so, whether they are unique in the sense of LO-uniqueness. These questions will be considered now.



Path	OD-pair	O	D	Links
A_1	A	I	III	1
A_2	A	I	III	2,3
B	B	I	II	2
C	C	II	III	3

Figure 4. The network considered to study existence and uniqueness of internal second-best optima

Specifically, a small example network, displayed in Figure 4, is used to demonstrate numerically that both questions can *not* in general be answered affirmatively. The network has three links (1, 2, 3), three nodes (I, II, III), and three OD-pairs: A (I-III), B (I-II), and C (II-III). For OD-pair A, two paths A_1 and A_2 are available; the other two OD-pairs each have only one path (B and C). N denotes use, and subscripts denote OD-pairs, paths or links. It is assumed that each link has a linear average cost function (a parameter's superscript denotes whether it represents the intercept (i) or slope (s) of a linear function):

$$c_1(N_{A_1}) = c_1^i + c_1^s \cdot N_{A_1} \tag{24}$$

$$c_2(N_{A_2} + N_B) = c_2^i + c_2^s \cdot (N_{A_2} + N_B) \tag{25}$$

$$c_3(N_{A_2} + N_C) = c_3^i + c_3^s \cdot (N_{A_2} + N_C) \tag{26}$$

For OD-pairs A and C, linear inverse demand functions apply:

$$D_A(N_{A_1} + N_{A_2}) = D_A^i - D_A^s \cdot \left(N_{A_1} + N_{A_2}\right) \tag{27}$$

$$D_C(N_C) = D_C^i - D_C^s \cdot \left(N_C\right) \tag{28}$$

Total benefits for these two OD-pairs are determined by the relevant area under the inverse demand curve. The demand for OD-pair B, N_B, is – for convenience – assumed to be perfectly inelastic, and the total benefits (not the costs!) for this OD-pair are a given constant TB_B.

When on this simple network only on link 3 a toll f_3 can be charged, cases can be constructed where multiple local second-best optima exist, or where an interior second-best optimum does not exist but a corner-solution instead prevails. To facilitate discussion later, it is helpful to first briefly consider the various effects that such a toll on link 3 may have. Given the inelasticity of demand for OD-pair B, there are three such effects. First, the toll directly affects the use and congestion externality on link 3 itself. Secondly, in doing so, it may lead to increased use and congestion on link 1, by diverting users for OD-pair A from path A_2 onto path A_1. Thirdly, to the extend that this happens, the users of OD-pair B may benefit from reduced congestion on link 2. The second-best optimal toll tries to balance these three effects.

In the event that path A2 is irrelevant in the second-best optimum, all second-best aspects pertaining to f_3 vanish, and the following tax rule f_3^0 can be derived:

$$f_3^0 = mec_3 \tag{29}$$

where mec_j is again used as a short-hand for the marginal external congestion costs on link j, $N_j \cdot c_j^s$. The second-best tax f_3^0 has a 'quasi first-best' structure, because no indirect effects are relevant in the second-best optimum: a marginal change in f_3^0 away from mec_3 would reduce the efficiency of use of link 3 by path C, without affecting the use of the other two links by the other paths.

When path A2 is relevant, in contrast, the following more complicated second-best tax rule f_3^1 can be derived after some tedious calculations:

$$f_3^1 = mec_3 + \frac{D_C^s \cdot \left(D_A^s \cdot \left(mec_1 - mec_2\right) + c_1^s \cdot mec_2\right)}{D_A^s \cdot \left(c_2^s - D_C^s\right) + c_1^s \cdot \left(D_A^s + D_C^s - c_2^s\right)} \tag{30}$$

It is the possibility of a 'regime shift' between path A_2 being relevant or not, inducing either f_3^0 or f_3^1 to be the relevant second-best toll, that may lead to the cases where either multiple local second-best optima exist, or where an interior second-best optimum does not exist.

5.1. Example 1: Multiple local second-best optima

The existence of multiple local second-best optima may in the present network occur when with a relatively high f_3, path A_2 is irrelevant and a relatively high f_3^0 applies; whereas with a relatively low f_3, path A_2 becomes relevant but requires a relatively low or even negative f_3^1. This could happen when link 1 is relatively heavily congested, so that the attraction of users of path A1 to path A2 creates a substantial benefit in terms of reducing congestion on link 1. The upper row in Table 1 shows the parameter values that were used to create this situation.[9]

[9] Because the sole purpose of the examples was to demonstrate the possibilities of multiple local optima (in the upper row of Table 1), and of the non-existence of an interior solution (in the bottom row), no attention was paid to representing any real-world situations with either set of parameters.

	D_A^i	D_A^s	D_C^i	D_C^s	N_B	TB_B	c_1^i	c_1^s	c_2^i	c_2^s	c_3^i	c_3^s
Multiple local optima	100	1	100	1	10	1200	0	10	40	0	50	0.035
No interior optimum	100	1	100	1	10	1200	62.5	0	0	1	50	0

Table 1. Parameters used to create multiple local optima and non-existence of an interior optimum

Under these conditions, two local welfare optima are found. This is shown in the left panel of Figure 5, depicting welfare as a function of f_3. The left local optimum involves a negative toll $f_3^1=f_3\approx-2.4$, which attracts users from the heavily congested link 1. The benefits of doing so more than compensate the efficiency losses on link 3 that result from increased use, which in turn partly results from induced additional demand from OD-pair C. To the right of $f_3\approx0.8$, however, f_3 has become so high that path A_2 becomes irrelevant. A local optimum is then found when the marginal external costs on link 3 are exactly internalized, for users from OD-pair C alone, at a toll level of $f_3^0=f_3\approx1.6$. The latter is also the global optimum, but an example where the left peak would be globally optimal can also easily be constructed.

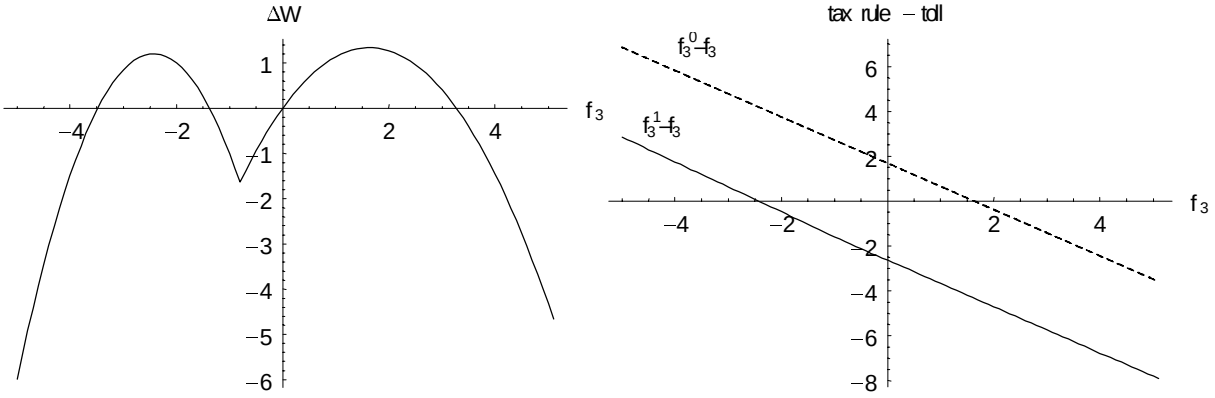


Figure 5. Welfare (left panel) and the difference between tax rules and the actual toll (right panel) with multiple local second-best optima

For each toll level f_3, the actual value of the tax rules f_3^0 and f_3^1 can be calculated according to equations (29) and (30), respectively. The right panel of Figure 5 shows the difference between the values thus obtained, and the toll f_3 actually charged. As expected, the curves intersect the horizontal axis at the respective local maxima, which indicates that tax rules (29) and (30) are indeed satisfied at $f_3^0=f_3\approx1.6$ and $f_3^1=f_3\approx-2.4$, respectively.

5.2. Example 2: Non-existence of an interior second-best optimum

The non-existence of an interior second-best optimum may in the present network occur when at a relatively low f_3, path A_2 is relevant and would require a relatively high f_3^1, whereas at a relatively high f_3, path A_2 is irrelevant and path C alone would require a relatively low f_3^0. This could be the case when link 2 is heavily congested while link 3 is hardly or not congested. The bottom row in Table 1 shows the parameters used to create such a situation.

The left panel of Figure 6 again depicts welfare as a function of f_3, and shows a maximum at $f_3=f^*\approx2.5$. The kink at f^* indicates already that this is not a 'standard' interior

optimum. To the right of this peak, path A2 is irrelevant, and f_3^0 would ideally exactly internalize the congestion externality on link 3 for path C alone. Given the absence of congestion on link 3, this would require a zero toll. To the right of f^*, welfare is therefore strictly decreasing in f_3. As soon as f_3 is smaller than f^*, however, path A2 becomes relevant, and demands a relatively high level of f_3^1, for the purpose of reducing congestion on the heavily congested link 2. Positive net benefits result from pricing off even the last A_2-driver on link 2, so that welfare is strictly increasing in f_3 to the left of f^*. This explains the kink at f^*.

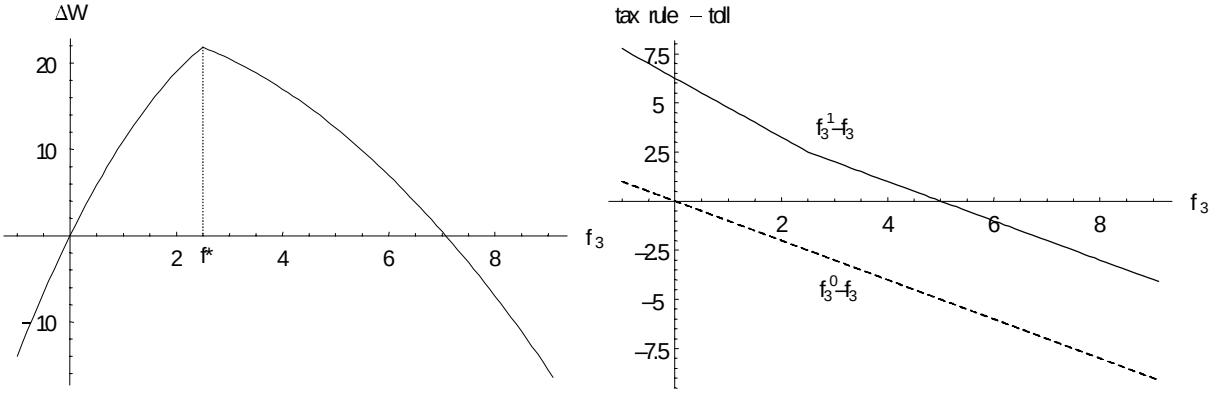


Figure 6. Welfare (left panel) and the difference between tax rules and the actual toll (right panel) with the non-existence of an interior second-best optimum

The right panel in the diagram further demonstrates that at $f_3=f^*$, neither (29) nor (30) is satisfied, so that the global optimum is not an interior solution for either tax rule. Instead, both rules are satisfied only for a value of f_3 outside the range for which the tax rule is relevant, at $f_3=0$ for f_3^0 and $f_3\approx5.0$ for f_3^1, which are given by the intersections with the horizontal axis.

5.3. Likeliness of occurrence

Having demonstrated the theoretical possibilities of the existence of multiple local second-best optima, and of the non-existence of an interior second-best optimum, a next question involves the likeliness of these complications occurring in practice. A general answer cannot be given, but some factors affecting these possibilities can be identified.

An important observation is that both examples require the existence of an untolled link that is significantly more heavily congested than the tolled link. Dependent on whether the use of this other link can be affected with the toll under control, its second-best optimal level is strongly affected by congestion on this other link. As a consequence, these examples involve situations where the regulator in fact has failed to choose the most attractive link to toll, under the restriction that only one toll can be set. These links would have been link 1 in the first example, and link 2 in the second (which cannot be seen from the diagrams). Apart from leading to a higher welfare level, interior second-best optima exist for these tolls, which is unique for Example 1. Example 2 has an infinite number of interior second-best optima for a toll on link 2, which is due solely to the fact that the demand by OD-pair B is assumed to be perfectly inelastic, and that path B uses link 2 alone and exclusively near the second-best optimum. With elastic demand, a unique interior optimum would have been found.

This suggests that the complications discussed may arise more easily, the less optimal the choice of the link that is subject to tolling. Specifically, a second-best toll will be less efficient when the various sub-goals it tries to achieve require more strongly conflicting incentives. This may induce relatively large changes in the optimal level once a certain path becomes relevant or irrelevant, which is what is driving the results in the two examples given.

Moreover, the examples involve rather extreme differences in the links' relative capacities and (hence) marginal external costs. When the slopes of the cost functions for the heavily congested links are reduced by a factor 10, and the intercepts are adjusted, kinks in the welfare as a function of f_3 are still observed, but unique interior second-best solutions are nevertheless found. This is shown in Figure 7, where the left panel shows the results corresponding to Example 1 (the parameters adjusted are c_1^i=80, c_1^s=1), and the right panel those corresponding to Example 2 (the parameters adjusted are: c_2^i=9, c_2^s=0.1). Also now, the tolls are still not levied on the most preferable link (which are the same as before), 'regime shifts' occur near the second-best optimal toll level, and there is still a significant variation in the slopes of the cost functions and marginal external costs across links. Nevertheless, a unique interior local and global second-best optimum is now found for both cases.

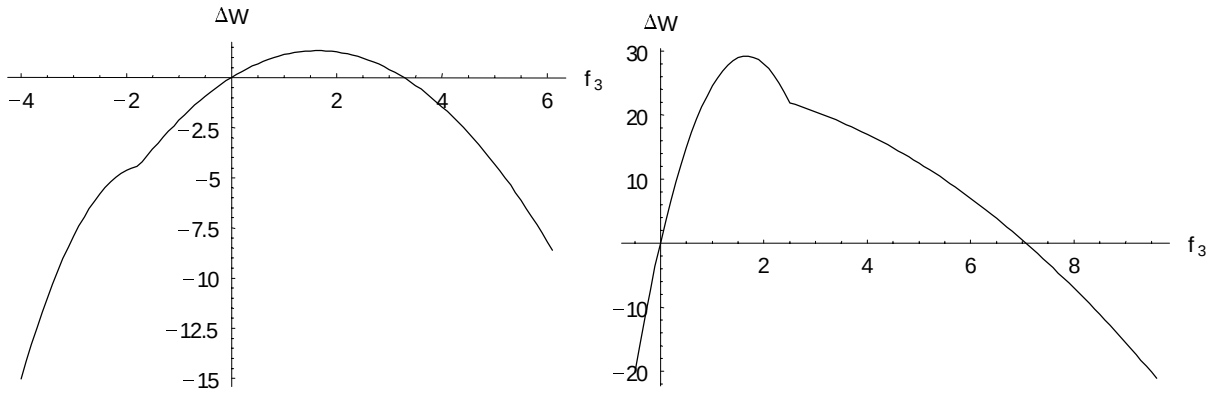


Figure 7. Welfare with a variation in parameters for Example 1 (left panel; c_1^i=80, c_1^s=1) and Example 2 (right panel; c_2^i=9, c_2^s=0.1)

In conclusion, it seems that non-uniqueness or non-existence of interior second-best optima are more likely to occur when marginal external congestion costs vary strongly over links, and when tolls are set on links that are not relatively heavily congested themselves, but that interact closely with links that are relatively heavily congested, where this interaction occurs through users of a path that may switch between relevance and irrelevance over a sufficiently small range of toll levels. The examples suggest that rather extreme conditions have to apply for non-uniqueness or non-existence to occur, implying that for practical applications, this need not be a major worry. Nevertheless, the theoretical possibility cannot be excluded. Finally, it is true that paths may be relevant or irrelevant depending on tolls set, so that welfare as a function of toll levels may show kinks. This fact alone, however, is not sufficient to lead to non-existence or non-uniqueness of interior second-best optima, as demonstrated in Figure 7.

6. Conclusion

This paper presented a general solution for the problem of second-best congestion tolling in static transportation networks where not all links can be tolled. With the existing plans for introducing electronic road pricing in many urban areas throughout the world, this type of problem is likely to become very important in the near future, in particular because it will often be considered inefficient (or unmanageable) to install the necessary equipment on all existing links. For small networks, the second-best tax rule may still yield analytically tractable congestion tolls, as was shown in the section where special cases of the general problem were discussed. Due to the occurrence of all sorts of cross-effects between tolled and untolled links, however, the tax rules will become solvable via numerical procedures only as the networks considered become more realistic and, as a consequence, larger. Nevertheless, the analysis presented has provided the necessary conditions for (interior) second-best optimality in such large networks, that can directly be applied regardless the size and the shape of the network. Moreover, it was demonstrated that, for instance by using the concept of 'virtual links', the analysis can even be applied rather easily to different classes of second-best problems in static networks as well.

In solving the general problem, an important set of variables used were the Lagrangian multipliers representing the 'shadow price of non-optimal pricing' for tolled and untolled links. The latter only play an indirect role in the second-best tax rule, reflecting indirect spill-over effects and interdependencies in networks that ought to be considered in second-best regulation. The former (the multipliers for the tolled links) are used directly in the optimization, in the sense that the absolute value of their sum is minimized. It was shown that in the first-best solution, these multipliers will all, individually, be equal to zero.

It was argued that the application of the first-order conditions and second-best tax rules obtained also for untolled links may provide guidelines for selecting that particular link for which it is economically most beneficial to add the next toll-point. Especially for large networks, where it is computationally too demanding to calculate the exact impact of adding a toll on each of the as yet untolled links, such procedures may be helpful.

Finally, a small example network was presented to demonstrate that interior second-best optima need not always exist, and if they exist, need not always be unique. The examples suggest, however, that for these complications to occur, rather extreme conditions have to apply in terms of the variation of marginal external costs over links. In addition, the regulator has to select relatively unattractive links to toll. This means that for practical applications, these need not be major worries. Nevertheless, the examples demonstrated that the theoretical possibility cannot be excluded.

References

Arnott, R.J. (1979) "Unpriced transport congestion" *Journal of Economic Theory* **21** 294-316.

Arnott, R., A. de Palma and R. Lindsey (1990) "Economics of a bottleneck" *Journal of Urban Economics* **27** 11-30.

Braid, R.M. (1989) "Uniform versus peak-load pricing of a bottleneck with elastic demand" *Journal of Urban Economics* **26** 320-327.

Braid, R.M. (1996) "Peak-load pricing of a transportation route with an unpriced substitute" *Journal of Urban Economics* **40** (179-197).

Dafermos, S. (1973) "Toll patterns for multiclass-user transportation networks" *Transportation Science* **7** 211-223.

Dafermos, S. (1980) "Traffic equilibrium and variational inequalities" *Transportation Science* **14** 42-54.

De Palma, A. and Y. Nesterov (1998) "Optimization formulations and static equilibrium in congested transportation networks" Paper presented to the 8th WCTR-conference, 12–17 july 1998, Antwerp, Belgium.

Glazer, A. and E. Niskanen (1992) "Parking fees and congestion" *Regional Science and Urban Economics* **22** 123-132.

Hearn, D.W. and M.B. Yildirim (1999) "A toll pricing framework for traffic assignment problems with elastic demand". Research Report 99-8, Department of Industrial and Systems Engineering, University of Florida.

Lévy-Lambert, H. (1968) "Tarification des services à qualité variable: application aux péages de circulation" *Econometrica* **36** (3-4) 564-574.

Liu, N.L. and J.F. McDonald (1999) "Economic efficiency of second-best congestion pricing schemes in urban highway systems" *Transportation Research* **33B** 157-188.

Marchand, M. (1968) "A note on optimal tolls in an imperfect environment" *Econometrica* **36** (3-4) 575-581.

May, A.D. and D.S. Milne (2000) "Effects of alternative road pricing systems on network performance" *Transportation Research* **34A** 407-436.

Nagurney, A. (1999) *Network Economics: A Variational Inequality Approach* (revised second edition) Kluwer Academic Publishers, Dordrecht.

d'Ouville, E.L. and J.F. McDonald (1990) "Optimal road capacity with a suboptimal congestion toll" *Journal of Urban Economics* **28** 34-49.

Pigou, A.C. (1920) *Wealth and Welfare*. Macmillan, London.

Small, K.A. and J.A. Gomez-Ibañez (1998) "Road pricing for congestion management: the transition from theory to policy". In: K.J. Button and E.T. Verhoef (1998) *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* Edward Elgar, Cheltenham (1998).

Smith, M.J. (1979) "The marginal cost pricing of a transportation network" *Transportation Research* **13B** 237-242.

Sullivan, A.M. (1983) "Second-best policies for congestion externalities" *Journal of Urban Economics* **14** 105-123.

Verhoef, E.T. (1999) "Time, speeds, flows and densities in static models of road traffic congestion and congestion pricing" *Regional Science and Urban Economics* **29** 341-369.

Verhoef, E.T. (2000) "Second-best congestion pricing in general networks: algorithms for finding second-best optimal toll levels and toll points". Manuscript, Free University Amsterdam.

Verhoef, E.T., R.H.M. Emmerink, P. Nijkamp and P. Rietveld (1996) "Information provision, flat- and fine congestion tolling and the efficiency of road usage" *Regional Science and Urban Economics* **26** 505-529.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1996) "Second-best congestion pricing: the case of an untolled alternative" *Journal of Urban Economics* **40** (3) 279-302.

Verhoef, E.T. and K.A. Small (1999) "Product differentiation on roads: second-best congestion pricing with heterogeneity under public and private ownership" Discussion paper TI 99-066/3, Tinbergen Institute, Amsterdam-Rotterdam.

Vickrey, W.S. (1969) "Congestion theory and transport investment" *American Economic Review* **59** (Papers and Proceedings) 251-260.

Wardrop, J. (1952) "Some theoretical aspects of road traffic research" *Proceedings of the Institute of Civil Engineers* **1** (2) 325-378.

Wilson, J.D. (1983) "Optimal road capacity in the presence of unpriced congestion" *Journal of Urban Economics* **13** 337-357.

Yan, H. and W.H.K. Lam (1996) "Optimal road tolls under conditions of queueing and congestion" *Transportation Research* **30A** (5) 319-332.

Appendix: Relaxing the assumption of link-specific congestion

The assumption of strictly link-specific congestion, that was made for the general model presented in the main text, may become problematic if *intersections* are to be modelled more realistically. The formulation used in the main text could only be applied directly to intersections if all users of an intersection to the same extent suffer from, and contribute to congestion on that intersection. In that case, the intersection could of course be treated as another link, just like the other links. In more realistic formulations, however, the representation with only link-specific congestion can in fact be considered as too restrictive. This can be illustrated by considering the intersection depicted in Figure A.1, where two two-way streets intersect. The dashed links, showing the 12 possible ways of using the intersection, cross each other in various cases, and hence direct congestion cost interdependencies between links are very likely to exist.[10] Moreover, the size of the (marginal) cost interdependency certainly needs not be constant over all pairs of links on the intersection. For instance, two different 'turns-to-the-right' will hardly hinder each other, while 'turns-to-the-left' and 'straight-ons' will generally be more conflicting (note that it is assumed that drivers use the right side of the road).

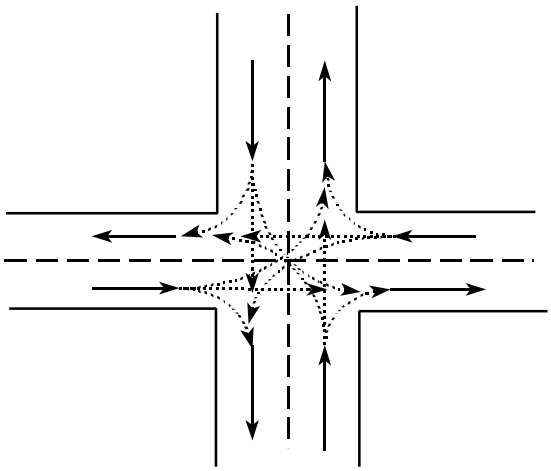


Figure A.1 Direct cost interdependencies between links on a simple intersection

Fortunately, it is rather straightforward to incorporate the implied direct cost interdependencies in the main model presented in equations (6)–(11). Doing so of course further complicates the analysis and the various expressions, but the main conclusions remain similar. We therefore present the equivalent expressions (A6)–(A11) below, for the more general case where the average user costs on link x may possibly depend on the level of usage N_j on all other links j, without further detailed comments.

First, the Lagrangian (6) can now be written as:

[10] It is very important to distinguish between these *direct cost interdependencies* between links, and the *indirect cost interdependencies* between links, that were mentioned also in the main text. Direct cost interdependencies result from the technical, direct interactions between users from different links; indirect cost interdependencies are caused by the equilibrating behaviour of users as described in Wardrop's first principle, leading to equalized equilibrium cost levels for all used routes between given OD-pairs.

$$\Lambda = \sum_{i=1}^{I} \int_{0}^{\sum_{p=1}^{P} \delta_{ip} \cdot N_p} D_i(x_i)dx_i - \sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{p=1}^{P} \delta_{jp} \cdot \delta_{ip} \cdot N_p \cdot c_j\left(\sum_{k=1}^{I}\sum_{q=1}^{P} \delta_{xq} \cdot \delta_{kq} \cdot N_q \,\forall x = 1,...,J\right)$$
$$+ \sum_{i=1}^{I}\sum_{p=1}^{P} \delta_{ip} \cdot \lambda_p \cdot \left[\sum_{j=1}^{J} \delta_{jp} \cdot \left(c_j\left(\sum_{k=1}^{I}\sum_{q=1}^{P} \delta_{xq} \cdot \delta_{kq} \cdot N_q \,\forall x = 1,...,J\right) + \delta_j \cdot f_j\right) - D_i\left(\sum_{q=1}^{P} \delta_{iq} \cdot N_q\right)\right] \tag{A6}$$

Note that the only difference between (6) and (A6) is that in (A6), c_j possibly depends on the link-flow on all links in the network.

The following necessary first-order conditions can now be derived:

$$\frac{\partial \Lambda}{\partial N_p} = \sum_{i=1}^{I} \delta_{ip} \cdot D_i - \sum_{j=1}^{J} \delta_{jp} \cdot \left(c_j + \sum_{x=1}^{J}\sum_{k=1}^{I}\sum_{q=1}^{P} \delta_{xq} \cdot \delta_{kq} \cdot N_q \cdot \frac{\partial c_x}{\partial N_j}\right)$$
$$+ \sum_{k=1}^{I}\sum_{q=1}^{P} \delta_{kq} \cdot \lambda_q \cdot \left(\sum_{j=1}^{J} \delta_{jp} \cdot \sum_{x=1}^{J} \delta_{xq} \cdot \frac{\partial c_x}{\partial N_j}\right) - \sum_{i=1}^{I}\sum_{q=1}^{P} \delta_{ip} \cdot \delta_{iq} \cdot \lambda_p \cdot D_i' = 0 \quad \forall \; p \text{ with } \delta_{ip} = 1 \tag{A7}$$

$$\frac{\partial \Lambda}{\partial f_j} = \sum_{i=1}^{I}\sum_{p=1}^{P} \delta_{ip} \cdot \delta_{jp} \cdot \lambda_p = 0 \quad \forall \; j \text{ with } \delta_j = 1 \tag{A8}$$

$$\frac{\partial \Lambda}{\partial \lambda_p} = \sum_{j=1}^{J} \delta_{jp} \cdot \left(c_j + \delta_j \cdot f_j\right) - \sum_{i=1}^{I} \delta_{ip} \cdot D_i = 0 \quad \forall \; p \text{ with } \delta_{ip} = 1 \tag{A9}$$

Note that only (A7) has changed, reflecting that the marginal external costs of a trip may now concern more links, namely also those links x for which the cross effect $\partial c_x/\partial N_j$ is positive for any link j which is part of the path p. The second line of (A7) shows that, for the same reason, a larger number of other path-flows may now be affected by marginal changes in N_p.

Substitution of (A9) into (A7) for each p for which $\delta_{ip}=1$ subsequently yields the following expression for the Lagrangian multipliers λ_p:

$$\lambda_p = \frac{\sum_{j=1}^{J} \delta_{jp} \cdot \left(\sum_{x=1}^{J}\sum_{q=1}^{P} \delta_{xq} \cdot N_q \cdot \frac{\partial c_x}{\partial N_j}\right) - \sum_{q=1,q\neq p}^{P} \lambda_q \cdot \left(\sum_{j=1}^{J} \delta_{jp} \cdot \sum_{x=1}^{J} \delta_{xq} \cdot \frac{\partial c_x}{\partial N_j}\right) + \sum_{i=1}^{I}\sum_{q=1,q\neq p}^{P} \delta_{ip} \cdot \delta_{iq} \cdot \lambda_q \cdot D_i' - \sum_{j=1}^{J} \delta_{jp} \cdot \delta_j \cdot f_j}{\sum_{j=1}^{J} \delta_{jp} \cdot \sum_{x=1}^{J} \delta_{xp} \cdot \frac{\partial c_x}{\partial N_j} - \sum_{i=1}^{I} \delta_{ip} \cdot D_i'} \tag{A10}$$

$\forall \; p \text{ with } \delta_{ip} = 1 \quad \text{and} \quad \forall \; q \text{ with } \delta_{iq} = 1$

As in the model in the main text, also here the system of equations (A10) should in principle have a unique solution for each λ_p, because it again makes up a system of X equations, generally linearly independent, in X unkowns (the λ_p's), where X denotes the number of relevant paths in the second-best optimum.

Finally, substitution of (A10) into (A8) gives the following expression for the second-best optimal congestion fees:

$$f_j = \frac{\sum_{p=1}^{P} \delta_{jp} \cdot \dfrac{\sum_{m=1}^{J} \delta_{mp} \cdot \left(\sum_{x=1}^{J} \sum_{q=1}^{P} \delta_{xq} \cdot N_q \cdot \dfrac{\partial c_x}{\partial N_m} \right) - \sum_{q=1, q \neq p}^{P} \lambda_q \cdot \left(\sum_{m=1}^{J} \delta_{mp} \cdot \sum_{x=1}^{J} \delta_{xq} \cdot \dfrac{\partial c_x}{\partial N_m} \right) + \sum_{i=1}^{I} \sum_{q=1, q \neq p}^{P} \delta_{ip} \cdot \delta_{iq} \cdot \lambda_q \cdot D_i' - \sum_{m=1, m \neq j}^{J} \delta_{mp} \cdot \delta_m \cdot f_m}{\sum_{m=1}^{J} \delta_{mp} \cdot \sum_{x=1}^{J} \delta_{xp} \cdot \dfrac{\partial c_x}{\partial N_m} - \sum_{i=1}^{I} \delta_{ip} \cdot D_i'}}{\sum_{p=1}^{P} \dfrac{\delta_{jp}}{\sum_{m=1}^{J} \delta_{mp} \cdot \sum_{x=1}^{J} \delta_{xp} \cdot \dfrac{\partial c_x}{\partial N_m} - \sum_{i=1}^{I} \delta_{ip} \cdot D_i'}} \qquad \text{(A11)}$$

$\forall\ j$ with $\delta_j = 1$ and $\forall\ p$ with $\delta_{ip} = 1$ and $\forall\ q$ with $\delta_{iq} = 1$

which is again quite similar to (11).

Indeed, in general, the extension presented in this appendix is, from the analytical viewpoint, only a minor one. The interpretation of the model's solution as given in (A10) and (A11) is largely analogous to the interpretation of the model with only link-specific congestion, with the main differences being the generally increased number of path-flows that are affected by each individual path-flow, and the fact that the congestion effects $\partial c_x / \partial N_j$ need of course not be equal for each j (for a given x), whereas in the simpler model only terms c'_j played a role.

However, notwithstanding the modest extension from an analytical viewpoint, the generalized results given in this appendix makes the model presented in the main text of course applicable to an even wider range of network problems, which justifies the discussion just given.